DOCUMENT RESUME

ED 242 760 TM 840 173

AUTHOR McKinley, Robert L.; Reckase, Mark D.
TITLE Implementing an Adaptive Testing Program in an Instructional Programs Environment.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Apr 84
CONTRACT N00014-82-K0716
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Adaptive Testing; Adults; *Computer Assisted Testing; *Instructional Materials; Item Banks; Military Training; Program Implementation; *Testing Problems; Testing Programs
IDENTIFIERS *Great Lakes Naval Training Center IL

ABSTRACT

The purpose of this paper is to identify and discuss some of the problems presented by the use of computerized adaptive testing (CAT) in an instructional programs environment versus large scale testing applications, and to describe an actual implementation of CAT in an instructional programs setting. This particular application is in the Electronic Technicians "A" (ETA) school at the Great Lakes Naval Training Center, Illinois. The goals of implementing CAT at this site were to increase test security, improve the efficiency of the testing program, and improve the quality of measurement yielded by the testing program. The problems encountered by this CAT program include the unknown dimensionality of the tests, the small number of available items for the item pools, and the availability of item response data only for small samples. The overall design of the project includes four phases: (1) preliminary analyses and software design; (2) implementation of a computer-administered conventional test; (3) implementation of a dual testing program (conventional and CAT); and (4) elimination of the conventional testing program. If the results are positive, this project will demonstrate that adaptive testing can effect improvement in classroom testing. (Author/BW)

Implementing an Adaptive Testing Program in an Instructional Programs Environment

Robert L. McKinley

and

Mark D. Reckase

The American College Testing Program

The use of computerized adaptive testing (CAT) in an instructional programs environment presents a number of problems not encountered in large scale adaptive ability testing applications. Among these are problems due to the achievement nature of the tests employed. Additional problems arise due to the small scale and classroom orientation of the instructional programs. The purpose of this paper is to identify and discuss some of these problems, and to describe an actual implementation of CAT in an instructional programs setting. The discussion will begin with a description of the environment in which the CAT program was implemented, and a discussion of the special problems encountered. This will be followed with a description of the approach taken in this particular application. Finally, the implications of this project and its results will be briefly discussed.

The Implementation

Environment

This particular application of CAT is in progress at the Great Lakes Naval Training Center (GLNTC) at Great Lakes, Illinois. More specifically, the instructional program involved is in the Electronic Technicians "A" (ETA) school. It involves a six week course on radar that covers three major areas: primary power distribution, transmitters, and receivers. Each area is covered by a test given at the end of instruction on that area. Approximately 700 students take the radar course each year, though the exact number varies from year to year.

Students are separated into classes varying in size, but ranging around an average of about twenty. Classes are 'lock-stepped', rather than using individualized instruction, but not all classes are at the same point in the program at any one time. That is, at any one time some classes will be on the power section of the course, some will be covering transmitters, and some will be studying receivers. To further confuse matters, there are three shifts per day: a day shift, an evening shift, and a midnight shift. Thus, instruction and testing continue throughout a given twenty-four hour period. Moreover, not all classes within a shift are covering the same material.

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1984. This research was supported by Contract No. N00014-82-K0716 with the Personnel and Training Research Programs of the Office of Naval Research.

The conventional test covering primary power distribution had forty multiple-choice items, each having four choices. The transmitter and receiver tests each had thirty items. For each test there were two forms, with no items in common to the two forms. Thus, there were eighty items available for the power test, and sixty items available for each of the other two tests.

Goals of CAT at the GLNTC

The overriding concern for the testing program at the GLNTC is security. For various reasons, test security becomes compromised at a phenomenonal rate at this site, and the ETA school is no exception. Conventional paper-and-pencil tests become obsolete due to compromised security almost as soon as they are produced. Because of this, one of the major motivations for implementing CAT is the improvement of test security.

Another goal of CAT at the GLNTC is the improvement of the efficiency of the testing program. It is hoped that the implementation of CAT will significantly reduce the amount of time required for testing, as well as the amount of time required of the staff for test administration. A relatively large number of students must pass through the testing program in a relatively short time, and any improvement in efficiency will be very important.

Another important goal of the CAT program at the GLNTC is the improvement of the quality of measurement yielded by the testing program. Under the circumstances prevailing at the GLNTC, decision errors due to poor measurement can have serious consequences. Very little in the way of resources is available for remediation, for instance. It is very important, then, that examinees passed on to the next unit of study actually be competent on the preceding units. This is especially important when one considers that these students will eventually graduate and move on to the fleet, where they will be responsible for maintaining and operating ships' equipment. One would like to have some confidence that the people graduating from the ETA school have, indeed, mastered the material taught there.

Special Problems

While this section addresses directly problems encountered at the GLNTC, it is likely that many of these problems are typical of instructional programs elsewhere. Most of these problems are inherent to classroom achievement testing, rather than being due to any special circumstances unique to the GLNTC.

One of the most serious problems encountered in adaptive achievement testing centers around the dimensionality of the tests. Achievement tests tend to be constructed using a table of specifications covering a variety of topics. Such tests often are highly multidimensional. CAT, on the other hand, is typically based on models and procedures requiring the assumption of unidimensionality. The conventional GLNTC tests were based on tables of specification, so at the outset of the project the dimensionality of the tests was unknown.

Another problem encountered at the GLNTC stemmed from the fact that the conventional tests used were relatively short. No resources were available for a large item development project, so the CAT item pools had to be constructed from the items available from the conventional tests. Unfortunately, not many items were available, so the resulting item pools were rather small.

To further complicate matters, item response data for use in item calibration were available only for small samples. In large scale testing programs, data collection for item calibration is relatively simple. In classroom testing, however, it is difficult and time consuming to amass large sample sizes. This is made even more difficult by the great haste with which tests become compromised at the GLNTC.

There are many other problems encountered in adaptive achievement testing that must be considered when implementing a CAT program in an instructional programs environment. Among these are questions about the concurrent, predictive, and content validity of adaptive tests; the stability of achievement test dimensionality; and, the effects of computerized adaptive administration on item characteristics. All of these must be addressed if CAT is to be used in instructional programs settings.

The Approach Taken at the GLNTC

Overview

The overall design of the project includes four phases. The first phase involves preliminary analyses to aid in the design of the software, along with the actual designing of the software. The second phase of the project involves the implementation of a computer administered coventional test. The third phase includes the implementation of a dual testing system which includes both a computerized conventional testing program and a CAT program. The fourth phase involves elimination of the computerized conventional testing program and expansion to other areas. The project is currently in the second phase. Each of these four phases will now be discussed, and the outcomes of the completed phases will be presented.

Phase I

Three primary tasks were undertaken during the first phase of this project. The first task was the completion of a study using simulation data that was designed to compare two different calibration models under conditions believed to be similar to those which would be encountered at the GLNTC. The second task was the collection and analysis of response data for the conventional paper-and-pencil tests for use in selecting a calibration model to be used in conjunction with item pool building. The third task involved the design of the test administration software for adaptive testing, as well as software for a computer administered conventional test. All three of these tasks were addressed under the constraint that the computer hardware to be used had already been selected by others.

Task 1. For this task a two-stage study was conducted to compare the ability estimates yielded by adaptive testing procedures based on the one-parameter logistic (1PL) and three-parameter logistic (3PL) models. The first stage of the study employed real response data, while the second stage employed simulated response data.

In the first stage, response data for 3000 examinees were obtained for the forty item ACT Assessment Mathematics Usage subtest (The American College Testing Program, 1982). The first 2000 cases were used to obtain item parameter estimates for both the 1PL and the 3PL models using the LOGIST computer program (Wingersky, Barton, and Lord, 1982). Using these estimates, 1PL and 3PL adaptive tests were simulated using the response data for the remaining 1000 cases. Both adaptive testing procedures employed maximum likelihood ability estimation and maximum information item selection procedures. The two sets of ability estimates yielded by the two adaptive testing procedures were then compared.

In the second stage, response data for 3000 cases were generated according to the 3PL model using as true parameters the 3PL item parameter estimates from the first stage. True abilities were selected from the standard normal distribution. The first 2000 cases were used for 1PL and 3PL calibrations of the items, and the remaining 1000 cases were used to simulate 1PL and 3PL adaptive tests. The two sets of ability estimates yielded by the two adaptive testing procedures were compared to each other and to the true ability parameters.

Results of this study are reported in detail in McKinley and Reckase (1983). They are summarized in Table 1, which shows the intercorrelations of the ability estimates for the real data, and Table 2, which shows the intercorrelations for the simulation data. In general, the results of both stages of the study indicated that the 1PL and 3PL adaptive tests yielded very highly correlated ability estimates, and that there was no apparent advantage, in terms of ability estimation, to using one of the models over the other. This was attributed to the fact that, due to the small size of the item pool, both procedures administered a large proportion of the item pool to each examinee. Thus, the two procedures administered much the same set of items to each examinee, and therefore yielded much the same ability esimate for each examinee.

Task 2. The second task of Phase I involved the collection and analysis of response data for the items in the item pool using the conventional paper-and-pencil test forms. Analyses performed on these data include principal components analyses, item analyses, and calibrations for the 1PL and 3PL models. The goal of these analyses was the evaluation of the appropriateness of the 1PL and 3PL models (or any unidimensional item response theory model) for use with these data. Data were available for approximately 400-500 examinees.

Table 3 shows the item analysis (proportion-correct difficulties and point biserial discriminations) and IRT calibration results for the transmitter item pool. These data are similar to those obtained for the other

pools. The results of the item analyses and the item response theory (IRT) model calibrations shown in Table 3 indicated that the items were all quite easy. Proportion-correct scores below 0.5 were rare. Because of this, considerable difficulty was encountered in the estimation of the guessing parameter. The LOGIST program tended to set the guessing value for most of the items equal to a constant value. This would seem to imply that a model with a constant guessing factor could be used with these data.

It was also discovered that the item discrimination values varied considerably. In the study described under the proceding section, item discriminations were uniformly high. Due to this and the smallness of the item pool, the 1PL and 3PL adaptive testing procedures yielded similar results. For these data, the item pool was small, but items varied in discrimination. Therefore, it was unclear to what extent the above study would generalize to these data.

In order to investigate this, another simulation study was conducted. The sixty 3PL item parameter estimates obtained for the items on the transmitter test were used as true parameters. Using these, the simulation dat design employed under Task 1 above was again applied. Information cutoffs for the two procedures were selected to yield tests of roughly the same average length. Again, the 1PL and 3PL adaptive test ability estimates were compared to each other and to the known true abilities.

Table 4 shows the intercorrelation matrix for the 1PL and 3PL adaptive test ability estimates and the true abilities. As can be seen, the 3PL estimates had a slightly higher correlation with the true values than did the 1PL CAT estimates. Still, the 1PL and 3PL CAT estimates were highly correlated. The 1PL adaptive tests had an average test length of fifteen, while the average test length for the 3PL tests was thirteen. These results support the conclusion that the little there is to gain from use of the more complex 3PL procedure is probably not worth the added expense. Bear in mind that what advantages there are to the 3PL model come only with dramatically increased sample sizes, which in many cases might be impractical or impossible to obtain.

The results of the principal components analyses indicate that, while these tests are not truly unidimensional, there does tend to be a dominant first factor. The other factors present do not lend themselves to interpretation. They do not appear to be associated with content or item type, and are therefore probably not important.

Task 3. Based on the results of the first two tasks of this phase, the decision was made to base the adaptive testing system on the 1PL model. The procedure developed employs maximum likelihood ability estimation and maximum information item selection. Testing is terminated when no items remain unused that yield an item information value for the most recent estimate of ability greater than a specified minimum, or until twenty items have been administered. The examinee's ability estimate is increased by a fixed stepsize for a correct response and decreased by a fixed stepsize for an incorrect response until both a correct and an incorrect response have been obtained. Initial estimates of ability were selected so as to represent

difficulty values near the mode of the item pool information function. The actual values for these parameters will not be determined until the completion of Phase II is near.

In addition to the CAT software, a computerized conventional test administration program was produced during Phase I. This program administers to an examinee the same set of items as appeared on the paper-and-pencil form of the test. Items are administered in a randomized order for test security purposes.

Phase II

Three primary tasks are included in Phase II of this project. The first task is the implementation of a computerized conventional testing program. The second task is the collection and analysis of response data from the computerized conventional tests for the purpose of updating the calibration results for the CAT item pools. The third task involves research directed at the investigation of the effects of computer administration on item characteristics.

Task 1. Initiation of the computerized conventional testing program occurred in late February. The program was implemented simultaneously for the three areas - primary power distribution, transmitter, and receivers. As was indicated previously in this program, items are selected and administered in a randomized order.

The purposes of this program are twofold. First, the program is necessary for obtaining additional response data for the item pool calibrations that are not contaminated due to compromised test security. Second, this program will yield data useful for assessing the effects of computer administration on item characteristics, particularly item difficulty. To date, insufficient data have been collected for meaningful analysis.

Task 2. The second task of Phase II will include item analyses, IRT analyses, and factor analyses of both the paper-and-pencil data and the computerized testing data. The purpose of these analyses is to assess the effects of computer administration on item difficulty, item discrimination, and the dimensionality of the item pools. This phase will commence once sufficient data have been collected from the computerized conventional testing program.

Task 3. The nature of the third task of Phase II will depend on the results of the analyses of the data collected from the computerized conventional testing program. Once these data have been analyzed, it will be determined whether or not these new data can be combined with the old in order to obtain new item pool calibrations. If the two sets of data cannot be combined, adaptive testing will commence when sufficient data for calibration of the item pools have been obtained from the computerized conventional testing program.

Phase III

The primary tasks of Phase III include the initiation of the CAT program, and the evaluation of the validity of the CAT program. During this phase, the CAT and computerized conventional testing programs will be run concurrently. Each examinee will be administered both. The purpose of this is the collection of data useful for a direct comparison of the CAT program to the conventional testing program. Similarities in the results of the two types of test will be considered to be evidence in support of the validity of the CAT program.

Phase IV

The fourth phase of the project includes two main objectives. First, once sufficient evidence for the validity of the CAT program has been collected, the computerized conventional testing program will be eliminated. Also, at this point work will commence on the expansion of the project to include other courses in the ETA school, and perhaps to other schools.

Once the CAT program has replaced the conventional testing program, other, more long-term research projects will be undertaken. Among these are the investigation of the stability of the item pool dimensionality (and calibration results) over time. Also, research will be conducted on procedures for the calibration of new items for the CAT item pools.

Implications

This project is important far beyond any value assigned to the research results, which will be quite important in themselves. This added significance derives from the nature of the project itself - the application of adaptive testing in the classroom. If the results of this project are positive, it will demonstrate that adaptive testing can effect improvement in an area of great significance.

References

McKinley, R.L. and Reckase, M.D. (1983). An evaluation of one- and three-parameter logistic tailored testing procedures for use with small item pools (Research Report ONR83-1). Iowa City, IA: The American College Testing Program.

The ACT Assessment, Form 23B. (1982). Iowa City, IA: The American College Testing Program.

Wingersky, M.S., Barton, M.A., and Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Table 1

Intercorrelation Matrix for Ability Parameter Estimates for the Real Data

Ability		Adaptive Tests		Paper-and-Pencil Tests	
Estimate		1PL	3PL	1PL	3PL
Adaptive	1PL	1.00	0.77	0.89	0.87
	3PL		1.00	0.81	0.86
P & P	1PL			1.00	0.95
	3PL				1.00

Table 2

Intercorrelation Matrix for True and Estimated Abilities for the Simulation Data

Ability		True	Adaptive Tests		Paper-and Pencil Tests	
Estimate			1PL	3PL	1PL	3PL
True		1.00	0.88	0.82	0.90	0.89
Adaptive	1PL		1.00	0.81	0.93	0.92
	3PL			1.00	0.83	0.85
P & P	1PL				1.00	0.93
	3PL					1.00

Table 3

Item Analysis and IRT Calibration Results
for the Transmitter Item Pool

Item	Item Analysis		IRT Calibration			
	p	r	1PL b	3PL a	3PL b	3PL c
1	0.63	0.18	-0.84	0.34	-0.93	0.02
2	0.29	0.16	-1.21	0.29	2.08	0.02
3	0.67	0.23	-1.09	0.42	-1.05	0.02
4	0.94	0.54	-4.22	0.88	-2.70	0.02
5	0.85	0.14	-2.68	0.35	-3.23	0.02
6	0.87	0.44	-2.84	0.75	-1.94	0.02
7	0.89	0.24	-3.11	0.43	-3.19	0.02
8	0.94	0.12	-4.22	0.31	-5.71	0.02
9	0.66	0.32	-1.04	0.67	-0.70	0.02
10	0.79	0.07	-1.99	0.27	-2.94	0.02
11	0.89	0.41	-3.08	0.68	-2.25	0.02
12	0.83	0.37	-2.42	0.59	-1.92	0.02
13	0.93	0.46	-3.84	0.75	-2.66	0.02
14	0.52	0.24	-0.19	0.50	-0.06	0.02
15	0.97	0.72	-5.20	2.22	-3.01	0.02
16	0.54	0.24	-0.31	0.52	-0.16	0.02
17	0.95	0.22	-4.51	0.46	-4.39	0.02
18	0.98	0.17	-5.64	0.47	-5.48	0.02
19	0.77	0.34	-1.89	0.61	-1.45	0.02
20	0.73	0.31	-1.51	0.57	-1.19	0.02
21	0.51	0.24	-0.15	0.45	-0.01	0.02
22	0.85	0.14	-2.63	0.28	-3.80	0.02
23	0.57	0.27	-0.52	0.55	-0.35	0.02
24	0.95	0.61	-4.29	1.19	-2.35	0.02
25	0.86	0.38	-2.72	0.63	-2.07	0.02
26	0.96	0.20	-4.78	0.45	-4.79	0.02
27	0.88	0.42	-2.98	0.71	-2.10	0.02
28	0.83	0.37	-2.40	0.63	-1.82	0.02
29	0.96	0.67	-4.71	1.31	-2.58	0.02
30	0.87	0.51	-2.90	1.00	-1.66	0.02
31	0.96	0.00	-4.01	0.22	-7.97	0.21
32	0.83	0.08	-2.02	0.38	-2.00	0.21
33	1.00	0.04	-7.28	0.76	-4.99	0.21
34	0.56	0.23	-0.24	0.65	0.45	0.21
35	0.85	0.22	-2.20	0.75	-1.36	0.21
36	0.84	0.16	-2.13	0.65	-1.42	0.21
37	0.97	0.05	-4.36	0.44	-4.58	0.21
38	0.94	0.19	-3.40	0.84	-2.14	0.21
39	0.78	0.12	-1.62	0.50	-1.15	0.21
40	0.93	0.04	-3.36	0.27	-5.32	0.21
41	0.71	0.26	-1.10	0.87	-0.25	0.29
42	0.99	0.00	-5.41	0.27	-9.14	0.21

Table 3 (Continued)

Item Analysis and IRT Calibration Results
for the Transmitter Item Pool

Item	Item Analysis		IRT Calibration			
	p	r	1PL b	3PL a	3PL b	3PL c
43	0.88	0.19	-2.50	0.65	-1.76	0.21
44	0.90	0.26	-2.74	0.95	-1.57	0.21
45	0.97	0.00	-4.36	0.29	-6.67	0.21
46	0.88	0.08	-2.56	0.41	-2.56	0.21
47	0.91	0.07	-2.88	0.40	-3.03	0.21
48	0.90	0.12	-2.80	0.53	-2.33	0.21
49	0.97	0.15	-4.18	0.85	-2.66	0.21
50	0.98	0.05	-4.85	0.46	-4.92	0.21
51	0.78	0.12	-1.64	0.40	-1.40	0.21
52	0.87	0.14	-2.39	0.53	-1.93	0.21
53	0.85	0.14	-2.20	0.56	-1.65	0.21
54	0.96	0.03	-4.09	0.35	-5.17	0.21
55	0.78	0.11	-1.61	0.39	-1.36	0.21
56	0.99	0.13	-6.01	1.28	-2.99	0.21
57	0.89	0.11	-2.65	0.49	-2.33	0.21
58	0.24	0.15	1.75	0.46	25.37	0.21
59	0.85	0.17	-2.20	0.58	-1.62	0.21
60	0.89	0.23	-2.62	0.86	-1.58	0.21

Table 4

Intercorrelation Matrix for True and Estimated Abilities for Simulated Data for the Transmitter Item Pool

Ability Estimate		True	Adaptive	
			1PL	3PL
True		1.00	0.71	0.78
Adaptive	1PL		1.00	0.77
	3PL			1.00

Implementing an Adaptive Testing Program
in an Instructional Programs Environment

Abstract

The use of computerized adaptive testing (CAT) in an instructional programs environment presents a number of problems not encountered in large scale adaptive ability testing applications. Among these are problems due to the achievement nature of the tests employed. Additional problems arise due to the small scale and classroom orientation of the instructional programs. In this paper, some of these problems are identified and discussed. In addition, an actual implementation of CAT in an instructional programs setting is described, and the special problems encountered in that implementation are discussed.